Data Sanitization for Preventing Sensitive Rule Mining

Dayanand S. Patil[1] , Vishwanath D. Chavan[2]
Department of CSE
[1, 2] Walchand Institute of Technology, Solapur India

Abstract-*Data mining is the process of extracting the required or useful information from the large volume of data and the result of that is used for decision making. Several data mining algorithms have been developed.One of them is "Association Rule Mining".Association Rule Mining is the technique which shows the association among the attributes in database. We are using"Genetic Algorithm (GA)" which derive possible association from the frequent item set. Besides we have taken one more add-on algorithm that is "Apriori Algorithm" for finding Frequent Item Set.*

Keywords-Genetic Algorithm,Apriori Algorithm, Frequent Item Set, Association Rule, Data mining.

I. INTRODUCTION

Many of the malls, big bazaar generates the huge amount of data that contains the useful or confidential information which shows the association among the data items. For example the customers who are buying the bread will definitely buy the butter also so this shows the associvity among this two item set.

By using some data mining techniques the third party can extract the useful information that the malls and individuals do not want to disclose to the public. In this paper we are using Association Rule Mining technique for preventing this problem. Association rule mining is a technique which hides the sensitive rules by updating the original data set with a little effect on non sensitive patterns as possible. The rules which are already available in original data that is sensitive rules are called old rules and the rules which are generated in sanitized data by some modification in original rules are called new rules. The data of malls contains the association rules which show the associvity among data items so the third party will get the advantage of that. For hiding this associvity among the data items we are transforming the original data which shows the association among the data items into sanitized one by constructing the sanitization matrix which hides the association patterns among data items. Finally as an output we get the sanitized data.

II. BASIC CONCEPT

In our approach, a transaction database D is represented as a matrix inwhich the rows represent transactions and the columns represent the items. If D contains m transactions and n kinds of items, D is represented as an mXn matrix. The entry D(t,i) is set to 1 if item i is purchased by the customer in transaction t. Otherwise, it is set to 0.

e.g.:

	1	2	3		1	2	3
t1	1	0	1		1	0	0
t2	1	1	0	X	0	1	1
t3	0	0	1		1	1	0
t4	1	1	1		0	1	0

Original Matrix Sanitized Matrix

Here 1,2,3 are the items and t1,t2,t3,t4 are the transaction.

2.1.Terminology used in Apriori Algorithm

Support

Support is an indication of how frequently the itemset appears in the database.

e.g.	Items	Support
	{1}	100
	{1,2}	300
	{1,2,3}	200
	{1,3}	400
	{2,3}	200
	{2,1}	600

Frequent Item Set

It refers to a set of items that frequently appear together, for example, milk and bread.

Pruning

The pruning step removes the extensions of (k-1) item sets which are not found to be common, from being considered for counting support.

2.2. Terminology used in Genetic Algorithm

Population

It is a subset of all the possible solutions to the given problem.

Chromosomes

A chromosome is one such solution to the given problem.

Gene

A gene is one element position of a chromosome.

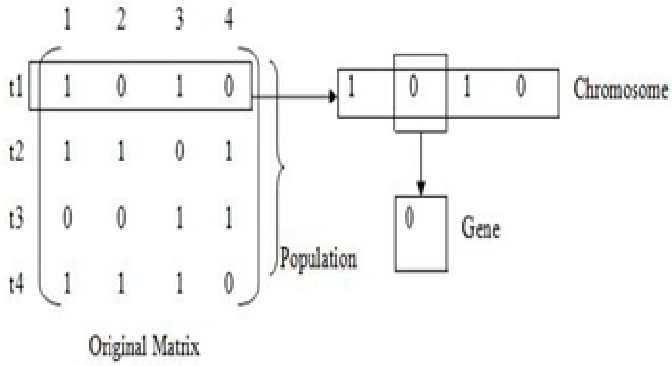


Fig: Terms used in Matrix

Fitness Function

A fitness function simply defined is a function which takes the solution as input and produces the suitability of the solution as the output.

III. BACKGROUND

3.1.Apriori Algorithm

Apriori algorithm is used to find frequent item set over transactional databases. These frequent item sets determined by Apriori can be used to determine association rules in the database.

In proposed approach Apriori uses support to derive frequent item sets. The item sets whose support is greater than or equal to specified support are used thus they are considered as frequent item sets and the item sets whose support is less than the specified support are pruned. To implement Apriori algorithm we have used MapReduce model and its pseudo-code is given below.

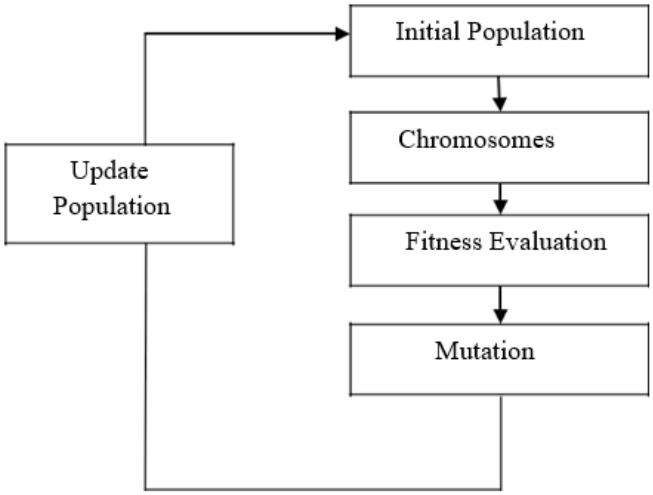
```
Map Task: // one for each split
Input: Si // Split i, line = transaction
Output: <key, 1> pairs, where key is an element
of candidate itemsets.
1. Foreach transaction t in Si
2.   Map(line offset,t)// Map function
3.    Foreach itemset I in t /* I = all possible sub
     sets of t */
4.      Out (I,1);
5.    End foreach
6.   End map
7. End foreach
8. End
```

```
Reduce Task:
Input: <key2, value2> pairs,
minimum_support_count, where key2 is an
element of the cand. Itemsets and value2 is its
occurrence in each split
Output: <key3, value3> pairs, key3 is an element
of frequent itemsets and value3 is its occurrence in
the whole dataset.
1. Reduce (key2, value2) // Reduce fun.
2.   Sum=0;
3.   While (value2.hasNext())
4.     Sum+= value2.getNext();
5.   End while
6.   If (sum>= min_sup_count)
7.     Out (key2, sum);
8. End if
9. End reduce
10. End
```

Figure: Pseudo-code of Apriori Algorithm

3.2. Genetic Algorithm

The genetic algorithm (GA) is an optimization and search technique based on the ethics of genetics and usual selection.In Genetic Algorithms, a population consists of a cluster of individuals called chromosomes that signify a complete solution to a certain problem. Each chromosome is a sequence of 0s or 1s.Each population consists of several chromosomes and the best chromosome is used to generate the next population. Based on the survival fitness, the population will make over into the future generation.



Initial Population

Population is a set of chromosomes. This population is initialized with random solutions.

Chromosomes

Chromosomes represents a solution to the problem which is composed of string of genes. The binary alphabet {0,1} is usually used to represent these genes.

Fitness Evaluation

In fitness evaluation,fitness is calculated for each chromosome using a fitness function.

Mutation

Mutation is a genetic operator which is used to maintain genetic diversity from one generation of population of chromosomes to the next generation.

3.3. Hadoop Framework

Hadoop is an open source framework which is used to process large data sets using programming models. Hadoop allows distributed storage as well as computation across cluster of computers that are built from commodity hardware. It consists of two parts storage part known as Hadoop Distributed File System(HDFS) and processing part called MapReduce programming.

The Hadoop Distributed File System(HDFS) provides a distributed file system that can be used to run on clusters consisting of number of small computers. In HDFS, the file to be processed is split into large blocks and these blocks are stored on number of computers in the cluster.

Using MapReduce model large amount of data can be processed in parallel. Two tasks performed in MapReduce are the map task and the reduce task. In map task the data is processed by creating several chunks of data whose output is further passed to reduce task. In reduce task shuffle and reduce operations are performed on data that is taken as a input from map task which after processing creates new output and stores it in HDFS.

IV. FUTURE SCOPE

There is a plan to compare the proposed algorithm with other existing algorithms if any to validate the robustness and effectiveness. In Future this tool can be extended to increase the efficiency and reduce the side effects by minimizing the modifications on database.

V. CONCLUSION

In this project, we have sanitized the transactional data using Apriori algorithm and genetic algorithm for hiding the sensitive rules. In genetic algorithm, a fitness function is used, based on this value the transactions are selected and the sensitive items of this transactions are modified with continuous evaluation operations .Thus, all the sensitive rules are hidden.

REFERENCES

[1] Guanling Lee, Chien-Yu Chang, ArbeeL.P Chen,"Hiding Sensitive Patterns in Association Rules Mining", Proceedings of the 28t Annual International Computer Software and Applications Conference (COMPSAC'04) 0730-3157/04© 2004 IEEE.

[2] Karla Taboada, Kaoru Shimada, Shingo Mabu, Kotaro Hirasawa and Jinglu Hu,"Association Rules Mining for Handling Continuous Attributes using Genetic Network Programming and Fuzzy Membership Functions", SICE Annual Conference 2007.

[3] S. Narmadha, S. Vijayarani, "Protecting Sensitive association Rules in Privacy Preserving Data Mining using Genetic Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 33– No.7, November 2011.

[4] Telikani, A. Shahbahrami and R. Tavoli,"Data sanitization in association rule mining based on impact factor", Journal of AI and Data Mining Vol 3, No 2, 2015.

[5] www.tutorialspoint.com